

# Prinzipien des Testens

## Grundlage:

Stochastisches Phänomen, über das eine Aussage gemacht werden soll.

## Beispiel:

Unterschied zwischen den Erwartungswerten  $\mu_1$  und  $\mu_2$  zweier Verteilungen

Beobachtung liefert unterschiedliche Mittelwerte

## Frage:

Sind die beobachteten Unterschiede zufällig und es gibt  $\mu_1 = \mu_2$  oder lassen diese den Schluss zu, dass  $\mu_1 \neq \mu_2$  gilt

# Illustration: Der gefälschte Würfel

**Gegeben:** Würfel, der möglicherweise gefälscht ist

**Experiment:** Würfeln

**Realisation:** mit R

60x Würfel fair →

Erwartete Häufigkeiten jeweils 10

Man beobachtet praktisch immer Abweichungen!

Wann lassen die Abweichungen auf einen systematischen Effekt schließen?

# Umsetzung in R

## Fairer Würfel

```
Würfel ← function (n) floor (runif (n, 1, 7));  
x=Würfel (60)  
table (x)
```

## Gefälschter Würfel

```
Würfel ← function (n)  
{hg = floor runif (n, 1, 7.2);  
  hg – (hg > 6)}
```

```
table (x)  
barplot (table (x))
```

Ausgabe der Daten  
Plot

# Statistische Inferenz

- Ist der beobachtete Unterschied zufällig oder steckt ein realer Effekt dahinter ?
- Lösung: Statistischer Test
- Ansatz: Wie könnten die Daten aussehen, wenn kein Unterschied vorhanden ist ?

# Hypothesen

Nullhypothese:

Kein Effekt (hier: Würfel fair, Mittelwerte gleich)

Alternative:

Es gibt einen Unterschied

# Zur Überprüfung des Testproblems

**Verdichtung** der Info aus Stichprobe in **Prüfgröße** bzw. **Teststatistik**

**Wichtig:**

Anhand Teststatistik Entscheidung darüber, ob eher  $H_0$  oder  $H_1$  **als allgemeine Aussage** zutrifft, d.h.  $H_0$  und  $H_1$

Hier: Prüfgröße: Betragmäßiger  
Unterschied  $T$  der Mittelwerte  
(normiert)

Falls  $T > c$

für geeigneten „kritischen“ Wert  $c$   
→ Entscheidung für  $H_1$ !

# Fehlentscheidungen

Test entscheidet

- Mittelwertsunterschied, obwohl dies nicht stimmt
- Kein Mittelwertsunterschied, obwohl ein Unterschied vorhanden ist

d.h.

- $H_0$  wird verworfen, obwohl  $H_0$  wahr  
→ Fehler 1. Art ( $\alpha$ -Fehler)
- $H_0$  wird beibehalten, obwohl  $H_1$  wahr  
→ Fehler 2. Art ( $\beta$ -Fehler)

# Fehler 1. Art und 2. Art

Damit sind folgende Ausgänge eines Tests möglich:

		Hypothese	
		wahr	nicht wahr
Test	lehnt ab	<b>Fehler 1. Art</b> ( $\alpha$ -Fehler)	richtig
	lehnt nicht ab	richtig	<b>Fehler 2. Art</b> ( $\beta$ -Fehler)

# Fehleranalyse

Wir können Fehlentscheidungen nicht ausschließen, aber es ist möglich, die **Wahrscheinlichkeit von Fehlentscheidungen** zu kontrollieren

- Konstruktion statistischer Tests so, dass **Kontrolle über** Wahrscheinlichkeit für **Fehler 1. Art** durch kleine **vorgegebene** obere Schranke
  - Signifikanzniveau  $\alpha$
  - Sicherheitswahrscheinlichkeit  $1 - \alpha$
- **keine Kontrolle über** Wahrscheinlichkeit für **Fehler 2. Art**
  - Suche nach **bestem** Test:  
unter allen Tests zum Niveau  $\alpha$  für vorliegendes Testproblem derjenige mit geringster Wahrscheinlichkeit für Fehler 2. Art

## Damit:

- Nullhypothese höchstens mit Wahrscheinlichkeit  $\alpha$  **fälschlicherweise** verworfen
- Wahrscheinlichkeit für den Fehler 2. Art **nicht vorgegeben**
  - abhängig von gewählter Alternative, **je näher** wahrer Parameter an (nicht wahren) Wert aus  $H_0$ , **desto größer** Wahrscheinlichkeit für Fehler 2. Art



**Ungleichbehandlung** beider Fehlerarten

- Grund für Formulierung eigentlicher Forschungsfrage als statistische Alternative: **Entscheidung für  $H_1$  durch  $\alpha$  statistisch abgesichert!**

# Herleitung der Verteilung

$$Y_{Ah} \sim N(\mu_1, \sigma^2)$$

$$\Rightarrow \bar{Y}_A \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$$

$$Y_B \sim N(\mu_2, \sigma^2) \Rightarrow \bar{Y}_B \sim N\left(\mu_2, \frac{\sigma^2}{m}\right)$$

$$\Rightarrow \bar{Y}_A - \bar{Y}_B \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

$$\Rightarrow T = \frac{\bar{Y}_A - \bar{Y}_B}{\sigma^2 \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Schätzung  $\sigma^2$   
aus empirischen Varianzen:

$$S_p^2 = \hat{\sigma}^2 = \frac{(n-1)S_{YA}^2 + (m-1)S_{YB}^2}{n+m-2}$$

$$S_{YA}^2 = \frac{1}{n-1} \sum_{l=1}^n (Y_{Al} - \bar{y}_A)^2$$

$$S_{YB}^2 = \frac{1}{m-1} \sum_{k=1}^m (Y_{Bk} - \bar{y}_B)^2$$

# Zwei-Stichproben t-Test

**Annahme:** Unabhängige normalverteilte Stichproben mit gleicher Varianz

$$Y_{Ak} \sim N(\mu_1, \sigma^2) \quad k = 1, \dots, n$$

$$Y_{Bl} \sim N(\mu_2, \sigma^2) \quad l = 1, \dots, m$$

**Hypothesen:**  $H_0 : \mu_1 = \mu_2$  vs.  $H_1 : \mu_1 \neq \mu_2$

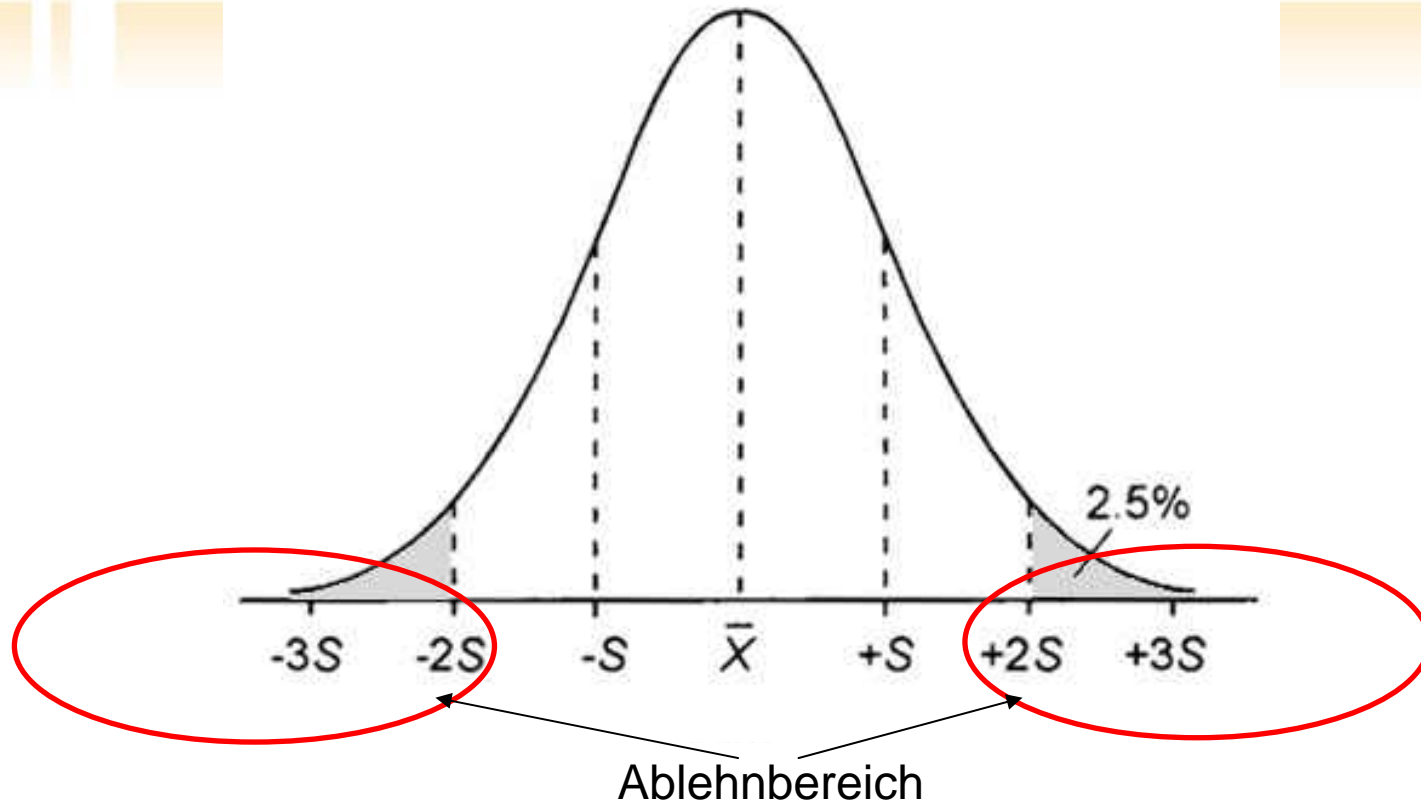
**Teststatistik:**

$$T = \frac{|\bar{y}_A - \bar{y}_B|}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$\text{mit } s_p^2 = \frac{(n-1)s_{yA}^2 + (m-1)s_{yB}^2}{n+m-2}$$

Lehne  $H_0$  ab, falls  $T > t_{1-\frac{\alpha}{2}}(n+m-2)$

# Finden des Ablehnungsbereiches



# P-Wert

Unter der Annahme, dass  $H_0$  wahr ist, bezeichnet man die Wahrscheinlichkeit, dass die Teststatistik den beobachteten Wert oder einen noch extremeren Wert (im Sinne von: „noch weiter weg von  $H_0$  liegend“) annimmt, als P-Wert des Signifikanztests. Je kleiner der P-Wert ausfällt, desto weniger passen die Nullhypothese  $H_0$  und die Stichprobendaten zusammen.

# P-Wert-Bestimmung

Versuch liefert Testgröße  $T_0$

$$p = P(T > T_0 | H_0)$$

Im Fall der Simulation der Würfel:  $P$  = Anteil der Werte unter  $T_0$

Im Fall des t-Tests

$$p = P(|T| > T_0 | H_0)$$

aus der t-Verteilung

# Signifikanz

- Statistische Signifikanz = Ergebnis nicht durch Zufall erklärbar
- Statistisch signifikant  $\neq$  Bedeutsam, wissenschaftlich relevant
- Daher immer Größe des Effekts (Schätzung) angeben !
- P-Wert kein direktes Maß für die Effektstärke
- Relevanz nur inhaltlich zu klären
- Vorsicht bei großen Stichprobenumfängen

# Theorie Statistischer Test

- Siehe Rüger: Test- und Schätztheorie Band II
- Neyman Pearson Theorie: Auffinden optimaler Tests unter gewissen Bedingungen
- Gütefunktion:  $G(\theta) = P(H_0 \text{ wird abgelehnt})$
- Niveau:  $G(\theta) \leq \alpha$  für  $\theta$  in  $H_0$
- Unverfälscht:  $\sup(H_0) G(\theta) \leq \inf(H_1) G(\theta)$
- UMP (Uniformly most powerful) : Gütefunktion dominiert jeden anderen Niveau-  $\alpha$  test im <Bereich der Alternative



# Überlegungen zu signifikanten Ergebnissen

Literatur Why Most Published Research Findings Are False  
John P. A. Ioannidis PLoS Med 2(8): e124.

Grundidee: Fehler bei Signifikanztests sind bedingte Wahrscheinlichkeiten unter  $H_0$  bzw  $H_1$ .

Die Berechnung der (marginalen) Wahrscheinlichkeiten ist abhängig von den a priori W'keiten von  $H_0$  bzw  $H_1$